BUSN 41201 - Big Data

Spring Quarter 2024 – Veronika Rockova

Integrating Machine Learning Techniques into the Hedonic Pricing Model for Enhanced Housing Price Prediction in Los Angeles County

May 26, 2024

Prepared by:

Will Sigal and Pablo Zavala

We pledge our honor that we have not violated the Honor Code during the preparation of this assignment

Table of Contents

0. Executive Summary

This paper analyzes the applications of modern statistical methods to housing price determination and prediction in Los Angeles County in 2021. We primarily focus on the influence of amenities' availability and neighborhood membership on these prices. The analysis starts with exploring the different variables gathered at the unit level from multiple governmental and private sources regarding the availability of amenities in Los Angeles in 2021, along with other relevant variables. We used OLS, Lasso, and spatial analysis methods to decompose prices into their most significant determinants and predictors. Finally, we composed a model to understand how some houses with similar features may display widely dissimilar prices. Our findings indicate that amenities, neighborhood characteristics, and spatial factors play crucial roles in determining housing prices, providing valuable insights for stakeholders in the real estate market. Our best model performance was with the Random Forest algorithm, which explained about 64% of the out-of-sample variance and achieved a mean squared error (MSE) of .147 when analyzing the logarithmic price per square foot of homes in Los Angeles County.

1. Introduction

Determining and predicting housing prices is a critical area of study in real estate economics, offering valuable insights for policymakers, real estate investors, and homebuyers. This paper delves into the application of modern statistical methods to understand the housing market dynamics in Los Angeles County in 2021, focusing on the influence of amenities and neighborhood characteristics on housing prices.

Previous literature extensively covers various factors influencing real estate prices,

utilizing methodologies like the Hedonic Price Method (HPM), which estimates the value of a property based on its characteristics and the amenities it offers. Studies have shown that neighborhood characteristics are often over-researched, while the implicit value of structural characteristics and social factors, such as crime rates, remain under-explored. This paper comprehensively analyzes both areas.

The research questions that guide this study are: How do various factors impact real estate prices in Los Angeles County? What are the key predictors of housing market trends in this region? To address these questions, we employed advanced statistical tools, including linear regression, Lasso regression, and cross-validation techniques, to identify determinants of housing prices and predict market trends.

Policymakers can leverage these findings to inform urban planning and housing policies, ensuring that developments meet market demands and address affordability issues. Real estate investors can use the insights to make informed decisions about property investments, understanding which factors most significantly drive property values. Homebuyers can better understand what contributes to housing prices, enabling them to make more informed purchasing decisions.

The data used in this study were gathered from multiple governmental and private sources, providing a robust foundation for analysis. The dataset includes unit-level variables such as price, square footage, number of bedrooms and bathrooms, presence of amenities like pools and garages, and locational attributes such as proximity to amenities and demographics.

Our analysis explores these variables through data visualization to uncover patterns and trends. Then, we apply relevant statistical models to decompose housing prices into their most significant determinants, highlighting how certain factors may contribute to price variations. In the advanced analysis section, we employ more sophisticated techniques to predict housing prices and understand broader real estate trends. Our findings may have significant implications for policy, investment, and personal decision-making, providing a solid foundation for future research and practical applications in the housing market in Los Angeles.

2. Data Collection and Preparation

Data Sources

Internal Factors:

The primary dataset we used was a housing dataset that included characteristics of all homes sold in California within the first six months of 2021. The dataset contains information on a home's location, price, living area, year built, bedrooms, bathrooms, whether it has a pool or a spa, the sale event (e.g., price reduction, sale, listed on the market), and posted time of the listing. The complete housing dataset had over 35,000 observations and gave us all the unit-level features we used for our HPM.

External Factors:

For external factors, we collected data on neighborhood features we found important during our literature review and through experience working in the real estate industry. The key metrics we used were area median income (census tract level), school concentration, crime rate,

population density, restaurant quality, green space near the home, and nearby foreclosures.

We used LA County's historical crime dataset for the crime rate, which included data on the occurrence date, type of crime, and location of all crimes reported in LA County between 2020 and 2021. This dataset had over three million observations. We also collected a list from LAHD of all property foreclosures from 2020 and 2021, which included the address and date of all 3100 foreclosures in Los Angeles during that period. To understand how proximity to schools might affect housing prices, we collected data on all of LA County's academic institutions and kept variables on their location, type (e.g., charter, private, public), and level (e.g., primary, secondary, college). To identify parks, we collected geospatial data on all parks and open spaces in LA County, which had information on a park's size and location. For restaurant quality, we

collected a list of the top 1,000 restaurants in LA created through the Yelp Fusion API. This dataset had the restaurant's name, address, review count, and rating. We used the 2020 5-year ACS estimates with the Census API for population and median income and put them all at the tract level.

Data Preprocessing and Cleaning

First, we subset our California housing data to just the area relevant to our research – LA County. With relatively uniform values, we put our dependent variable, price per square foot (PPSF), into a logarithmic scale. Then, we strung together all the address information before using the Google Places API to geocode every address and gave each home a unique geocode (latitudinal and longitudinal coordinates). We then deleted missing values and uniform valued variables and were left with 4,580 observations with 28 variables.

Compact Table of Housing Variables

Variable	Type
X	integer
is_forAuction	integer
event	character
price	numeric
pricePerSquareFoot	numeric
city	character
yearBuilt	integer
streetAddress	character
zipcode	numeric
longitude	numeric
latitude	numeric
livingArea	numeric
bathrooms	numeric
bedrooms	numeric
parking	integer
garageSpaces	numeric
hasGarage	integer
levels	character
pool	integer
spa	integer
hasPetsAllowed	integer
datePosted	Date
log_pricePerSquareFoot	numeric

Figure 1: Variables of housing before merging external factors

We called the ACS data for population and median income using the "tidycensus" package before merging them at the census tract level. Since census tracts generally have about 4,000 individuals in each tract, we created a density column for the number of people within each square mile. For our crime data, we subset it to crimes classified as violent or a felony before turning each occurrence into a

geopoint. After that, we used the "stjoin" function to aggregate the number of felonies or violent crimes reported within each census tract. We had about 500 missing census tracts in our crime data, which is not insignificant, and we chose to replace the values with our median value.

We then merged the census data with our housing data by using "stjoin" and matching each house observation with a census tract identified by the census multipolygon geographical object. For the restaurant, school, and foreclosure datasets, we used their geopoints to calculate how many of each are in a respective radial distance from a given house. For schools, we created a function to iterate and predict log PPSF and use the radial distance that minimized MSE, which was five miles. We also added an indicator variable of whether a private school, a proxy for local school quality, was within two miles of the housing unit. For restaurants, we counted the number of top restaurants within three miles of each home. Based on previous literature, we counted the number of foreclosures within one mile based on previous literature, which found that foreclosures' effect on price was maximized at one mile. To account for green space, we used our park

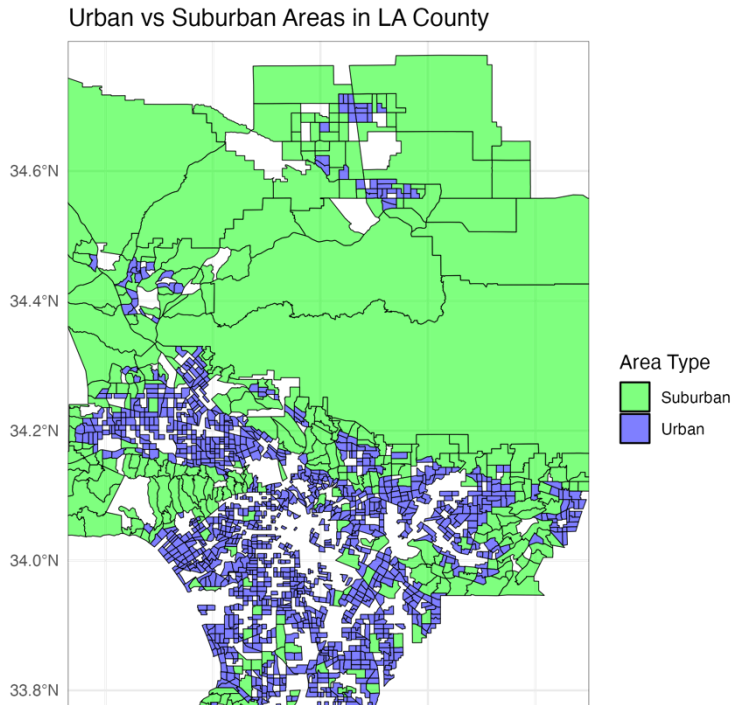


Figure 2: Map of the distributions of suburbs and urban areas in LA County

dataset and aggregated the acres of park and open space within two miles of each home. To account for the potential spatial biases of homes in the suburbs versus urban areas, we added an indicator variable to identify whether a home is in an urban area (see Figure 2) by analyzing whether the population density was higher than 5,000 people per square mile, which we chose due to its ability to identify the largest suburbs. Lastly, we found that in our initial lasso model, we underestimated homes located near the coast (see Figure 3), so we added an indicator variable or whether a home is located within half a mile from the coastline along with a continuous variable that gave the minimum distance in miles that each home is from the coastline.

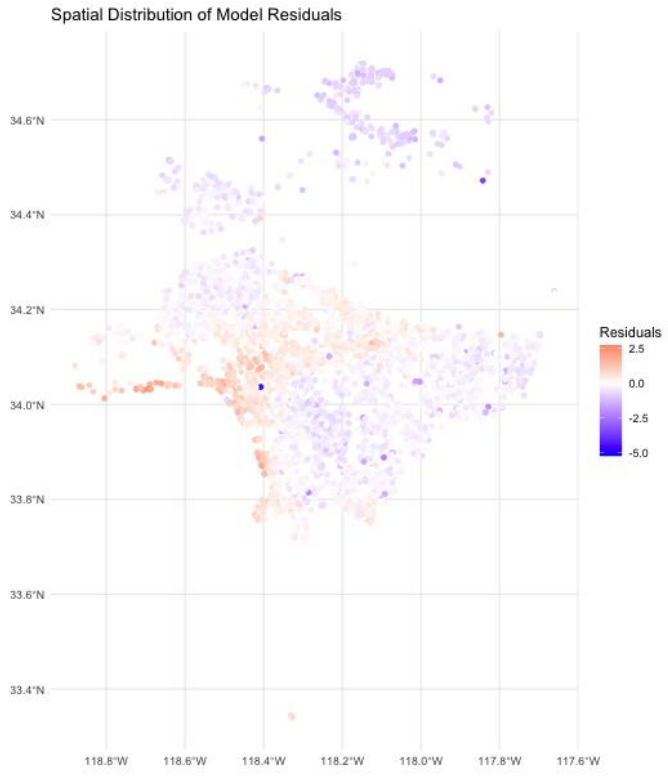


Figure 3: Spatial distribution of MSE from initial regression showing biases in Western suburbs and homes located near the coast.

3. Data Exploration and Visualization

Introduction to the Dataset

The merged dataset encompassed 3,803 housing unit-level observations and 38 variables, including transformed variables. These variables covered a wide range of attributes related to the housing units, such as physical characteristics (e.g., number of bedrooms, bathrooms, square footage), neighborhood characteristics (e.g., crime rate, median income), geographical data (e.g., latitude, longitude, neighborhood), and economic indicators (e.g., price, price per square foot). The dataset also included variables derived through data transformation and feature engineering aimed at enhancing the predictive power and interpretability of the models. Examples of such transformed variables include logarithmic transformations of price-related metrics to address skewness and improve linearity, which we mentioned in the previous section.

Summary Statistics

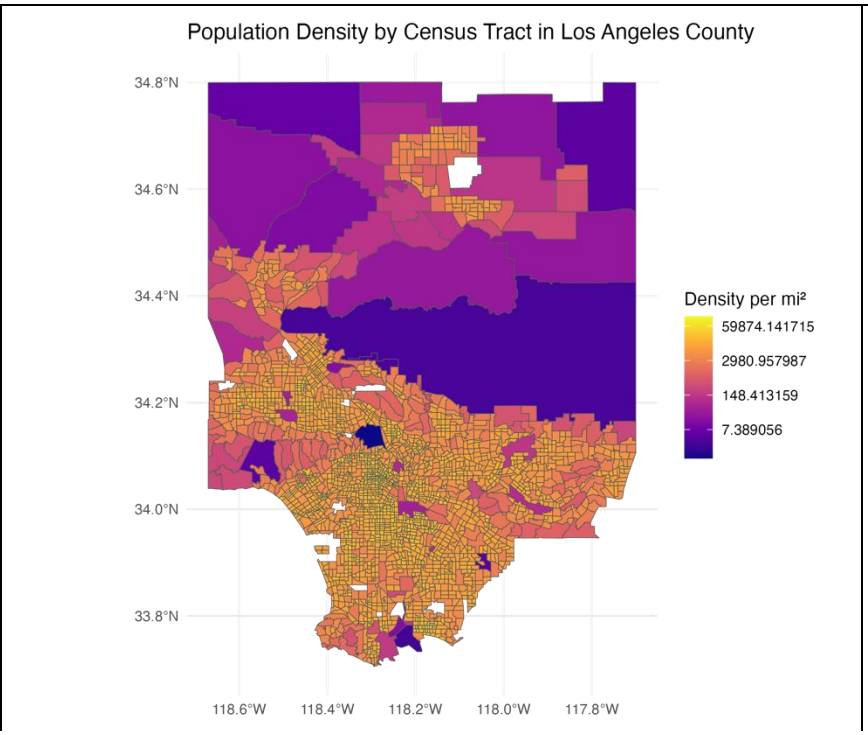
Our final merged dataset has both indicator, numeric, and categorical variables, which are summarized in the following table:

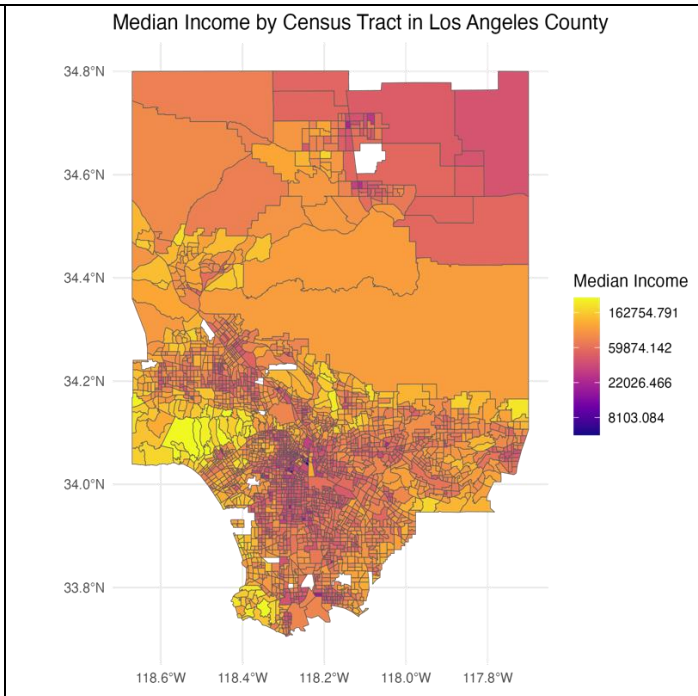
```
is_forAuction         event               city             yearBuilt      livingArea
Min.   :0.000000   Length:3803        Length:3803        Min.   :   0   Min.   :  378
1st Qu.:0.000000   Class :character   Class :character   1st Qu.:1942   1st Qu.: 1362
Median :0.000000   Mode  :character   Mode  :character   Median :1955   Median : 1845
Mean   :0.001052                                         Mean   :1955   Mean   : 2348
3rd Qu.:0.000000                                         3rd Qu.:1980   3rd Qu.: 2690
Max.   :1.000000                                         Max.   :2021   Max.   :41000
  bathrooms         bedrooms         parking        garageSpaces      hasGarage
Min.   : 0.000   Min.   : 0.000   Min.   :0.0000   Min.   : 0.000   Min.   :0.0000
1st Qu.: 2.000   1st Qu.: 3.000   1st Qu.:1.0000   1st Qu.: 1.000   1st Qu.:1.0000
Median : 2.000   Median : 3.000   Median :1.0000   Median : 2.000   Median :1.0000
Mean   : 2.872   Mean   : 3.641   Mean   :0.8012   Mean   : 1.649   Mean   :0.8028
3rd Qu.: 3.000   3rd Qu.: 4.000   3rd Qu.:1.0000   3rd Qu.: 2.000   3rd Qu.:1.0000
Max.   :25.000   Max.   :32.000   Max.   :1.0000   Max.   :18.000   Max.   :1.0000
   levels              pool             spa          hasPetsAllowed
Length:3803        Min.   :0.0000   Min.   :0.0000   Min.   :0.00000
Class :character   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.00000
Mode  :character   Median :0.0000   Median :0.0000   Median :0.00000
                   Mean   :0.2156   Mean   :0.1557   Mean   :0.01394
                   3rd Qu.:0.0000   3rd Qu.:0.0000   3rd Qu.:0.00000
                   Max.   :1.0000   Max.   :1.0000   Max.   :1.00000
log_pricePerSquareFoot  school_count   school_count_5miles   population   density_per_mi2
Min.   :1.099          Min.   :  0.0   Min.   :  0.0       Min.   :1045   Min.   :    2.4
1st Qu.:6.068          1st Qu.: 98.0   1st Qu.: 98.0       1st Qu.:3398   1st Qu.: 3931.5
Median :6.326          Median :154.0   Median :154.0       Median :4379   Median : 7253.6
Mean   :6.341          Mean   :166.1   Mean   :166.1       Mean   :4464   Mean   : 8399.0
3rd Qu.:6.617          3rd Qu.:210.5   3rd Qu.:210.5       3rd Qu.:5408   3rd Qu.:11031.6
Max.   :9.008          Max.   :501.0   Max.   :501.0       Max.   :9559   Max.   :70568.4
violent_crime_count violent_crime_per_person median_income    park_acres_within_2miles
Min.   :   1.0      Min.   :0.0001227        Min.   : 21296   Min.   :   0.00
1st Qu.:   4.0      1st Qu.:0.0010815        1st Qu.: 66636   1st Qu.:   0.00
Median : 139.0      Median :0.0326468        Median : 90741   Median :   0.00
Mean   : 169.8      Mean   :0.0384630        Mean   : 98305   Mean   : 155.26
3rd Qu.: 167.8      3rd Qu.:0.0450603        3rd Qu.:120379   3rd Qu.:  73.15
Max.   :2100.0      Max.   :0.5054859        Max.   :250001   Max.   :4929.68
top_1000_restaurants_3miles foreclosures_within_1mile miles_from_coast
Min.   :  0.0               Min.   : 0.000            Min.   : 0.02189
1st Qu.:  0.0               1st Qu.: 0.000            1st Qu.: 7.22271
Median :  0.0               Median : 0.000            Median :13.21069
Mean   : 26.3               Mean   : 6.537            Mean   :15.48754
3rd Qu.:  4.0               3rd Qu.:11.000            3rd Qu.:22.04359
Max.   :505.0               Max.   :66.000            Max.   :57.11875
half_a_mile_from_the_coast urban_suburban private_school_within_2miles
Min.   :0.0000             Min.   :0.00   Min.   :0.0000
1st Qu.:0.0000             1st Qu.:0.00   1st Qu.:1.0000
Median :0.0000             Median :1.00   Median :1.0000
Mean   :0.0263             Mean   :0.67   Mean   :0.9406
3rd Qu.:0.0000             3rd Qu.:1.00   3rd Qu.:1.0000
Max.   :1.0000             Max.   :1.00   Max.   :1.0000
```

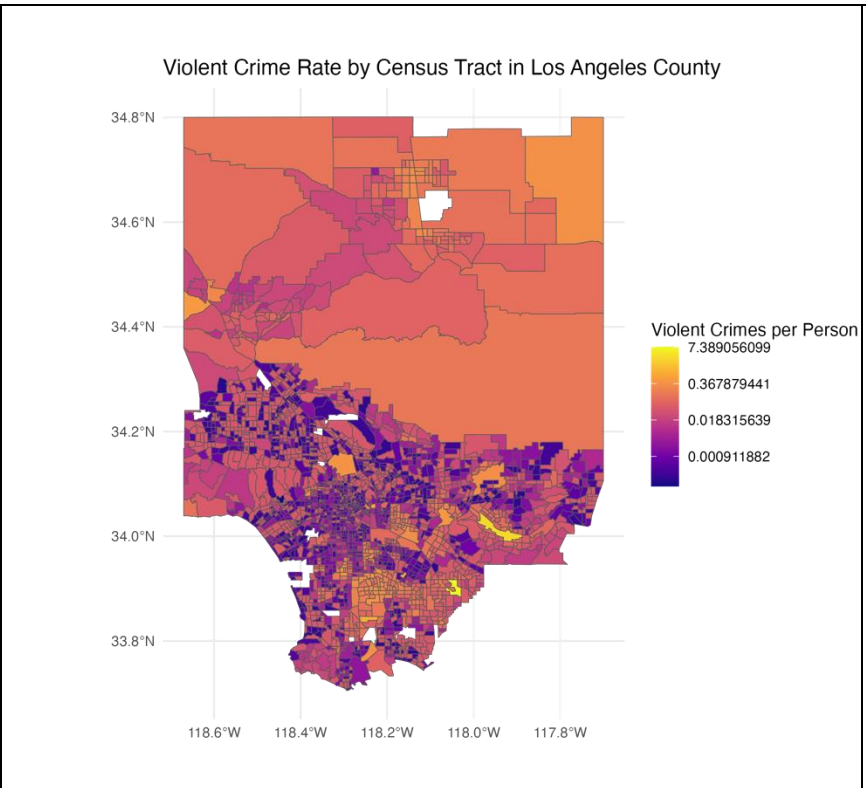
Figure 4: Summary of all our housing variables' distribution

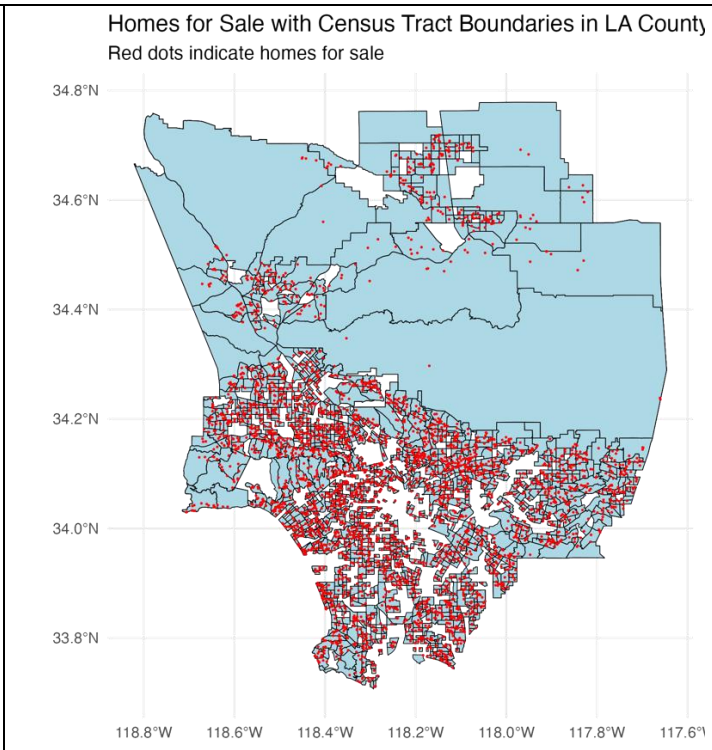
Our critical dependent variable, logPPSF, is normally distributed and has a median level of 6.326 with a standard deviation of .52. Other variables, like violent crime count (in each census tract), are much more variable with a median value of 169, and a standard deviation of 260.

Geospatial Distribution of Key Variables:









Analyzing these maps, we can see a few possible trends to look for in our later models. First, there seems to be a negative relationship between the areas with crime rates and median income. Furthermore, we see that few areas with high crime rates are on the coast but instead are primarily in areas in the south and relatively inland. With income, we can observe that high-income areas are in the farthest western point and are relatively commonly found along the coast. We can also observe that many of the most affluent areas have the lowest population density, indicating that the wealthiest neighborhoods are in the suburbs of Los Angeles. The last map shows the census tract of each of our homes. Since we use the merge "within" method in "stjoin,"

several census tracts are missing because there were no homes for sale in those tracts. Additionally, some census tracts only have a single home for sale, which could skew our data if the home for sale in that census tract is not representative of the typical home in that tract. However, sadly, this potential issue could not be avoided and is one of the limitations of our dataset.

Distribution of Housing Prices

Analyzing raw housing prices was inappropriate since the distribution exhibited skewness to the right due to extensive housing areas and costly square footage for some homes. This skewness can lead to misleading statistical conclusions and model predictions because most data points cluster around lower price ranges while a small number of high-priced homes disproportionately influence the results. To address this issue, we applied a logarithmic transformation to the housing price data, which helps to normalize the distribution and stabilize variance. This transformation makes the data more suitable for linear regression analysis by reducing the impact of outliers and making the relationship between variables more linear. We can visualize this problem and the effect of the transformation in the following histogram that plots prices against frequency:

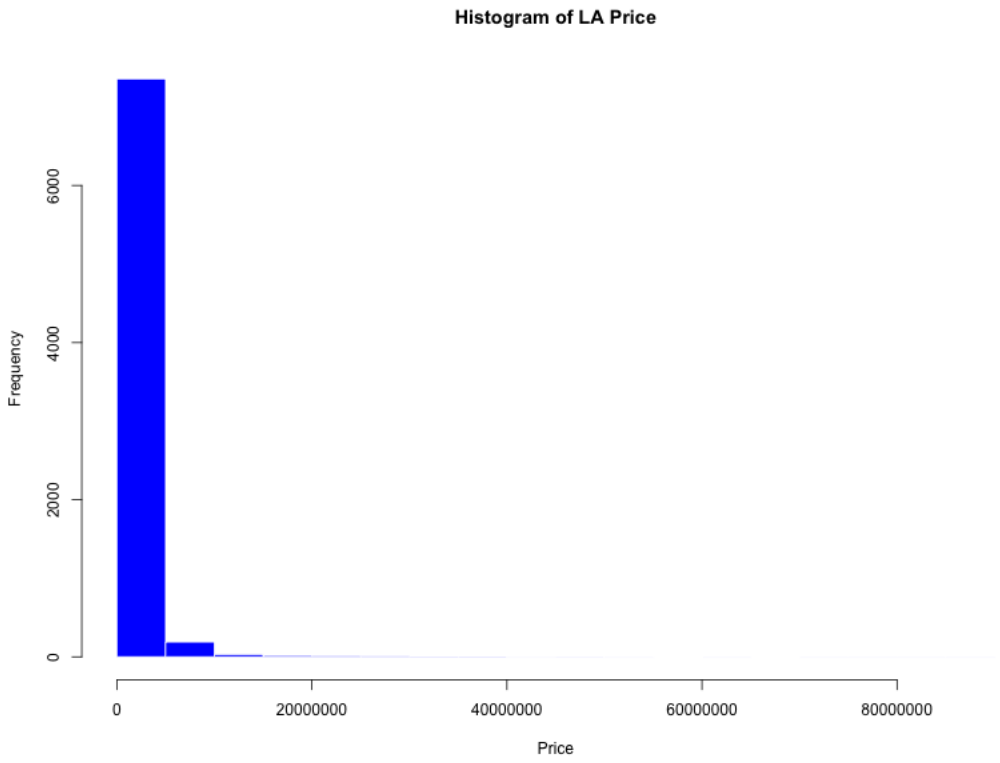


Figure 5: Spatial distribution of MSE from initial regression showing biases in Western suburbs and homes located near the coast.

Figure 5: Histogram of untransformed home price

The following histogram reflects the logarithmic price transformation per square foot, resulting in a more normalized data distribution.

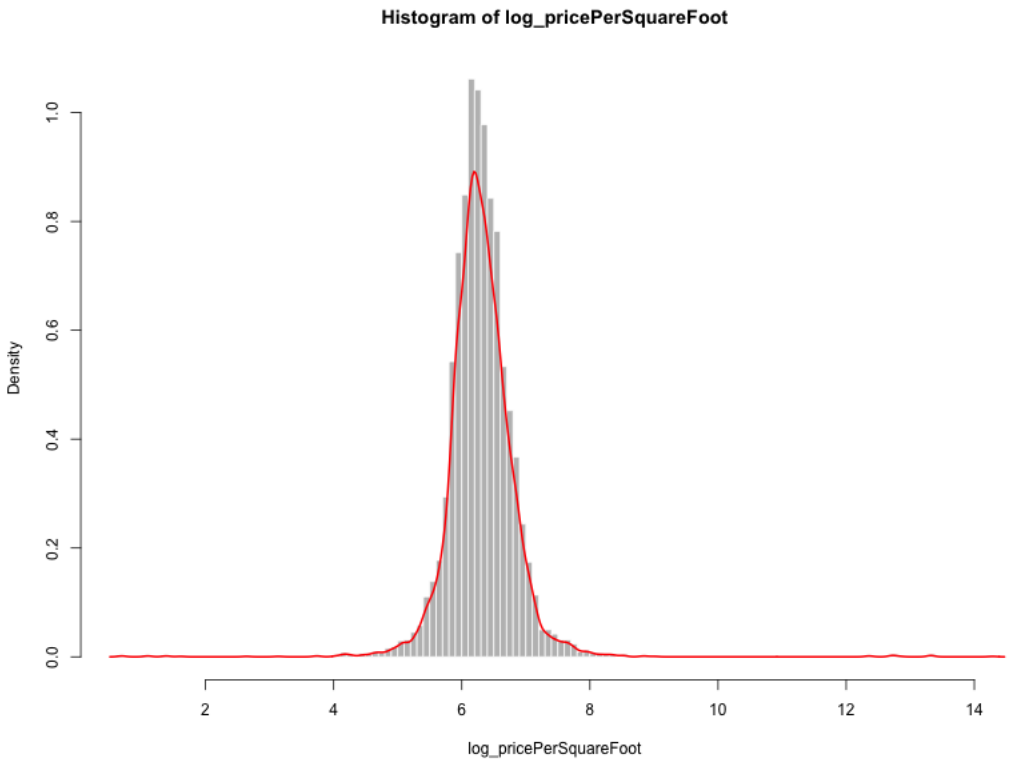


Figure 6: Histogram of the transformed home price

Correlation Analysis

A significant concern when analyzing features of the HPM is the presence of correlations between predictors, which become problematic when doing predictive or causal analysis due to multicollinearity. The summary of the correlations of internal factors is presented in the table below:

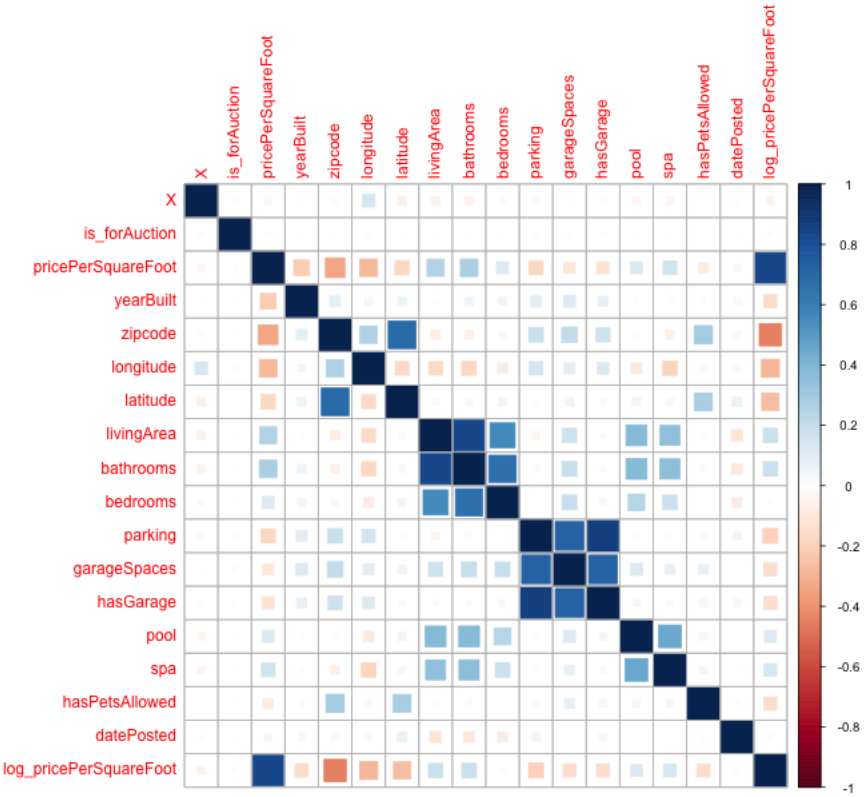


Figure 7: Correlation Matrix of our internal variablesFigure 6: Spatial distribution of MSE from initial regression showing biases in Western suburbs and homes located near the coast.

This correlation matrix visualizes the interrelationships between various housing-related internal factors. The correlation is positive for properties in the northern and western parts of the city, suggesting higher prices per square foot in these regions. There is a negative correlation with the year built, indicating that newer properties tend to have lower prices per square foot. Both positively correlate with price per square foot, which reflects that larger homes with more bathrooms command higher prices. Interestingly, there is a negative correlation between the

availability of parking and price per square foot, which may suggest higher value in more urban areas where parking is scarce, but properties are costly. The presence of a pool is positively correlated with price per square foot, underlining the added value of this amenity. These insights are critical for understanding how different characteristics influence regional housing prices.

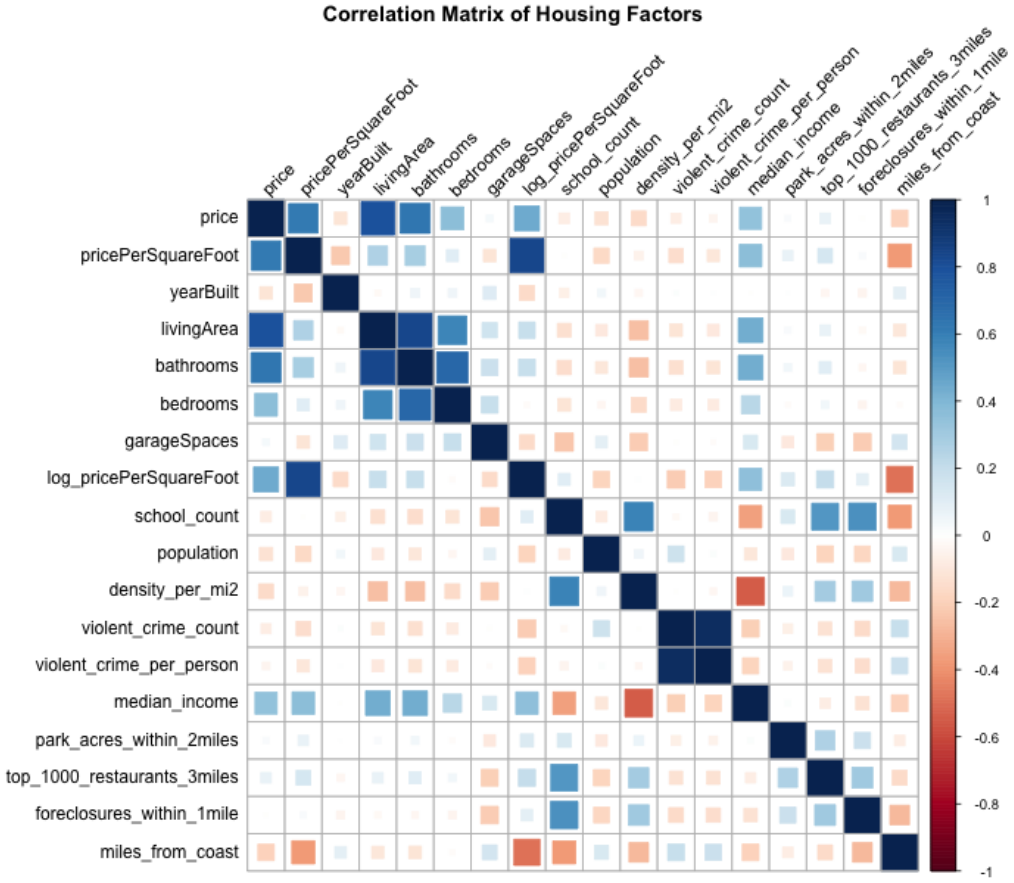


Figure 8: Correlation Matrix of our final dataframeFigure 7: Spatial distribution of MSE from initial regression showing biases in Western suburbs and homes located near the coast.

We can see many new insights by analyzing the correlation matrix with external

amenities included. Living areas exhibit a strong positive correlation with the number of bathrooms, bedrooms, and garage spaces, indicating that larger homes typically have more amenities and are often located in less spatially constrained areas.

The presence of schools within a given radius (school count) shows a positive correlation with log price per square foot, suggesting that proximity to schools positively influences property values. Population density (density per square mile) negatively correlates with price per square foot, indicating that higher population density areas might be associated with lower prices per square foot, possibly due to urban crowding. Violent crime total count and violent crime per person exhibit negative correlations with price per square foot and log price per square foot, highlighting that higher crime rates detract from property values.

Conversely, median income positively correlates with price per square foot, indicating that higher-income areas tend to have higher property values. As indicated by park acres within 2 miles, proximity to green spaces positively correlates with price per square foot and log price per square foot, underlining the added value of nearby parks. Similarly, proximity to quality dining options (top 1000 restaurants within 3 miles) correlates positively with price per square foot, suggesting that access to good restaurants enhances property values. Proximity to foreclosed properties (foreclosures within 1 mile) shows a negative correlation with price per square foot, indicating that nearby foreclosures can diminish home values. Lastly, distance from the coast negatively correlates with price per square foot and log price per square foot, showing that properties closer to the coast generally command higher prices. This correlation matrix is crucial

for identifying potential multicollinearity issues in the modeling phase and understanding how different factors interact and influence housing prices. The insights gained here will guide the selection of variables and the interpretation of subsequent regression models.

Clustering

A clearer understanding occurs when we classify our observations through clustering, an unsupervised machine-learning technique that groups data points into clusters based on the similarity of factors. We identified distinctive patterns and structures within the dataset by employing K-means clustering, allowing us to differentiate between housing units and neighborhoods based on their attributes. By analyzing these clusters, we can gain insights into how certain combinations of features impact housing prices and identify common characteristics shared by similarly priced homes. Using the K-means clustering algorithm, we partitioned our data into four relevant groups, which can be represented geographically in the following way:

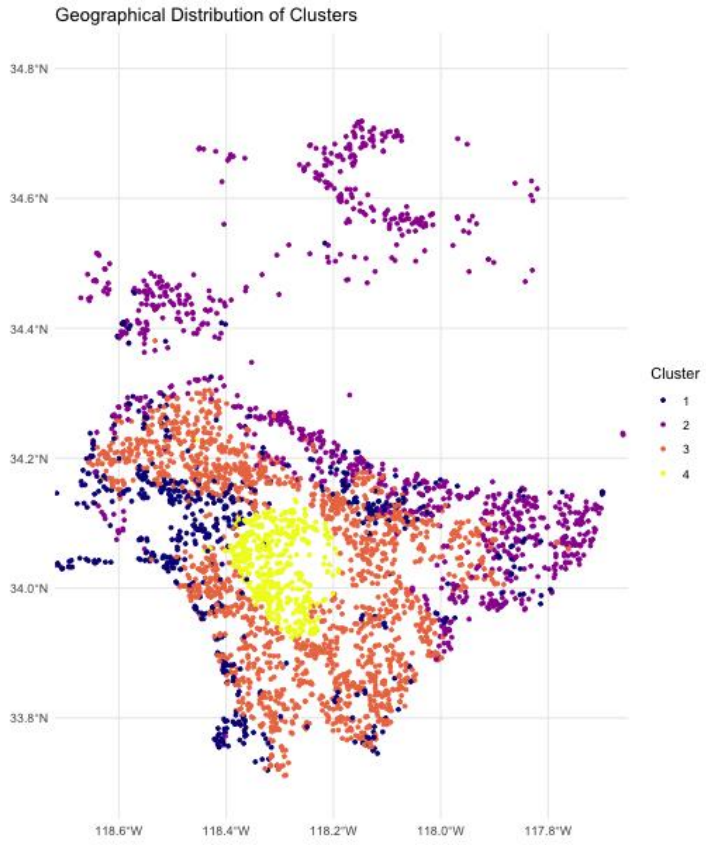


Figure 9: Geographical distribution of our K-means clusters

The clustering reflects a geographical separation, wherein housing groups pertain to a different cluster the further they are from the center of the city. The following are the clusters' characteristics:

```
##           X           yearBuilt    livingArea   bathrooms    bedrooms     garageSpaces
## 1 -0.18192084 -0.07716068  1.29015355  1.39545374  1.04936137    0.4953575
## 2  0.20207244  0.15402341 -0.14714798 -0.15271541 -0.03230037    0.2689418
## 3 -0.08446229 -0.02290147 -0.35873890 -0.39924598 -0.32282905   -0.1721203
## 4  0.16890718 -0.13265697 -0.09214632 -0.06762225 -0.14013676   -0.6130679
##   school_count school_count_5miles density_per_mi2 violent_crime_count
## 1   -0.4768576         -0.4768576      -0.6609249         -0.26020401
## 2   -0.9073844         -0.9073844      -0.6406474          0.16063393
```

```
## 3    0.1891692          0.1891692       0.3144622         0.06948625
## 4    1.9269588          1.9269588       1.0913158        -0.25909922
##   median_income park_acres_within_2miles top_1000_restaurants_3miles
## 1    1.25873714            -0.002598033                -0.2033556
## 2   -0.02091422            -0.197173776                -0.3256456
## 3   -0.28604459            -0.071454949                -0.2425924
## 4   -0.62472126             0.734286968                 2.0189339
##   foreclosures_within_1mile miles_from_coast
## 1               -0.11870299       -0.5025712
## 2               -0.54895723        1.1901169
## 3               -0.05007577       -0.3440715
## 4                1.59448033       -0.4736982
```

Based on the mean characteristic of each cluster, we can group our units with the following descriptions:

Cluster 1

New construction, larger living areas, and more bathrooms and bedrooms characterize this cluster. These housing units are closer to the coast, have higher median incomes, and have fewer foreclosures.

Cluster 2:

This cluster is characterized by relatively older homes with smaller living areas situated further from the coast. Higher violent crime rates and fewer amenities seem to be relevant features of these homes.

Cluster 3

Homes in this cluster are situated moderately close to some amenities, including schools and restaurants, and have mixed characteristics regarding size and age. There is a moderate crime rate and a balanced number of foreclosures.

Cluster 4

Cluster 4 features homes in densely populated areas with many schools and restaurants nearby. These homes have fewer garage spaces and are further from the coast but have higher park acreage within two miles.

It is helpful to observe the decision tree the algorithm follows to classify our observations to deepen our analysis of each cluster's characterization:

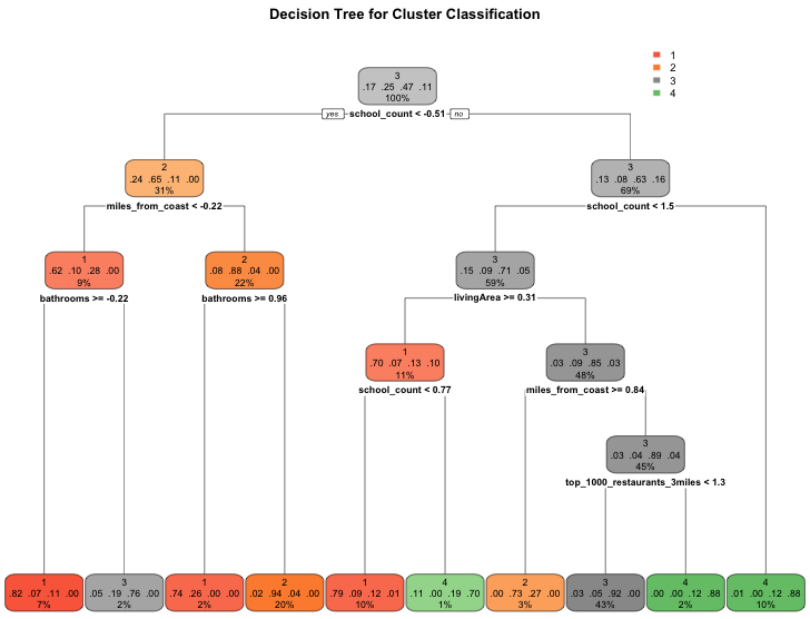


Figure 10: Decision Tree for our K-means clustering.

The decision tree segments the data based on the importance of specific features. At the root of the decision tree is the school_count feature, with a threshold value of 0.51. The first node of the tree indicates that the number of schools in the vicinity is a primary factor in determining the cluster classification.

Then, for clusters with fewer schools (school_count < 0.51), the proximity to the coast (miles_from_coast) becomes the next most significant factor. If a property is closer to the coast (miles_from_coast < -0.22), the number of bathrooms becomes the decisive factor. Properties with more bathrooms (bathrooms >= -0.22) are categorized into cluster 1, with more bathrooms, closer proximity to the coast, and fewer schools. Conversely, properties with fewer bathrooms (bathrooms < -0.22) fall into cluster 2.

The number of bathrooms continues to be significant for properties that are further from the coast (miles_from_coast >= -0.22). Properties with more bathrooms (bathrooms >= 0.96) are also categorized into Cluster 2, while those with fewer bathrooms fall into Cluster 3.

For clusters with more schools (school_count >= 0.51), the decision tree further splits based on the school_count < 1.5. Here, the livingArea becomes a critical factor. Properties with larger living areas (livingArea >= 0.31) are split based on their proximity to the coast and other amenities. Those closer to the coast (miles_from_coast >= 0.84) and with fewer top-rated restaurants nearby (top_1000_restaurants_3miles < 1.3) fall into cluster 3, indicating a higher density of amenities. Properties with smaller living areas fall into clusters 3 and 4 based on further splits in school count and other characteristics. However, when analyzing the variable importance for clustering classification:

```
##                      Overall
## bathrooms            788.6896
## bedrooms             232.5296
## density_per_mi2      665.0105
## livingArea          1361.9238
## median_income        379.1780
```

```
## miles_from_coast             836.0908
## park_acres_within_2miles     105.8101
## school_count                 931.2593
## school_count_5miles          931.2593
## top_1000_restaurants_3miles  506.3245
## X                            0.0000
## yearBuilt                    0.0000
## garageSpaces                 0.0000
## violent_crime_count          0.0000
## foreclosures_within_1mile    0.0000
```

Living area (1361.9238) is the most critical factor in cluster segmentation. Larger living areas generally correspond to higher housing prices and a better standard of living, making this feature highly influential. School Count and School Count within 5 Miles (931.2593 each) are also significant, highlighting the importance of educational facilities' proximity. Miles from the coast (836.0908) indicates that properties closer to the coastline tend to have higher values due to the desirable location and scenic views. The number of bathrooms (788.6896) is another critical variable, as more bathrooms usually indicate a larger, more luxurious home, contributing to higher property values. Density per Square Mile (665.0105) reflects the urbanization level of an area, with higher density often correlating with a bustling urban environment, various amenities, and higher property prices. Top 1000 Restaurants within 3 Miles (506.3245) shows that the availability of top-rated restaurants nearby affects housing prices by reflecting the quality of dining options and the overall lifestyle of the area. Median income (379.1780) indicates the economic status of a neighborhood, with higher median incomes correlating with more affluent communities and higher housing prices. Park Acres within 2 Miles (105.8101) is also a contributing factor, as access to green spaces and recreational areas enhances the desirability of a

neighborhood.

Variables such as Year Built, Garage Spaces, Violent Crime Count, and Foreclosures within 1 Mile have an importance score of zero, indicating they do not impact this clustering model. Distribution of housing prices serves to understand how these clusters differentiate from one another concerning logPPSF quartile belonging:

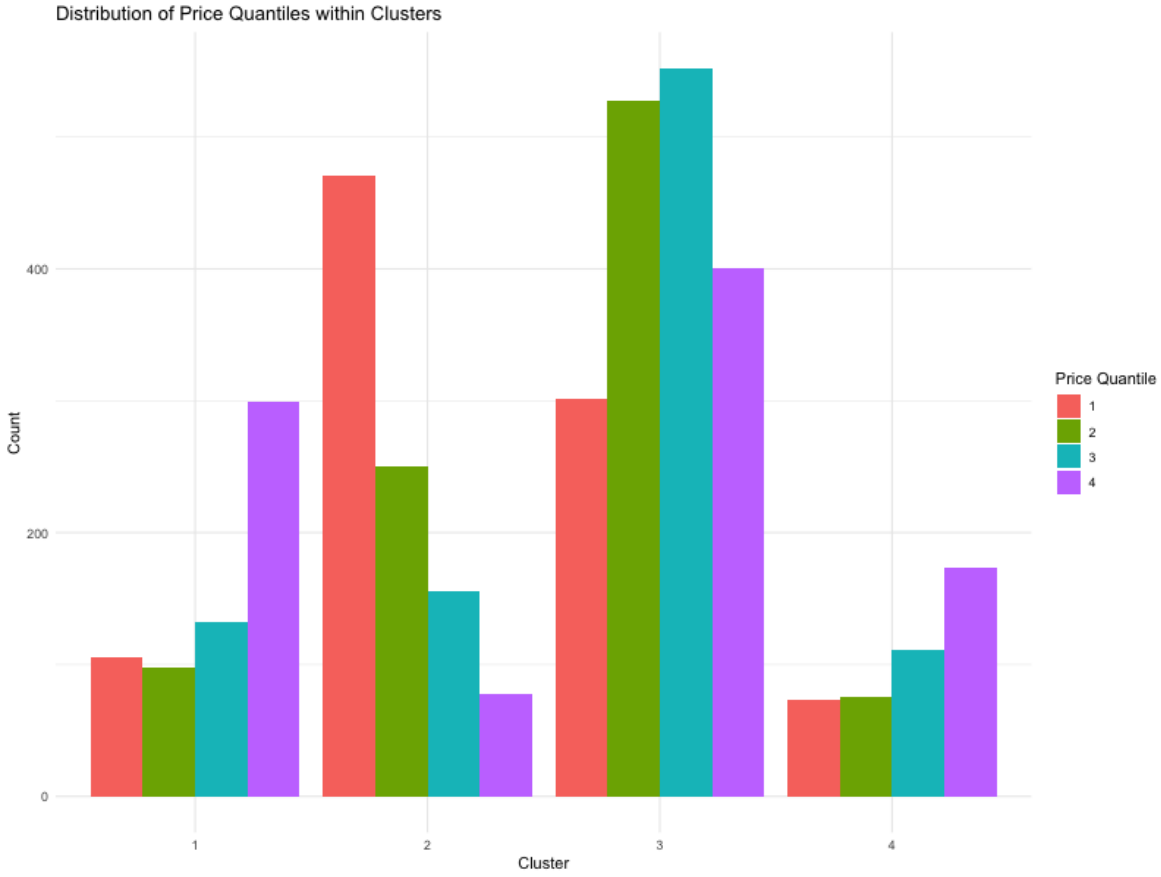


Figure 11: Frequency table showing price quantile distribution within each cluster.

Again, this graph is consistent with our previous analysis, allowing us to classify our housing

units into differentiated clusters effectively.

Principal Component Analysis

Following the correlation matrix we obtained, we create principal components that reduce the dimensionality of the data by grouping factors into the same PC, starting with the PC that explains most of the data's variance. When explaining the total variance of our housing price data using principal components, we obtain the following scree plot and variance table:

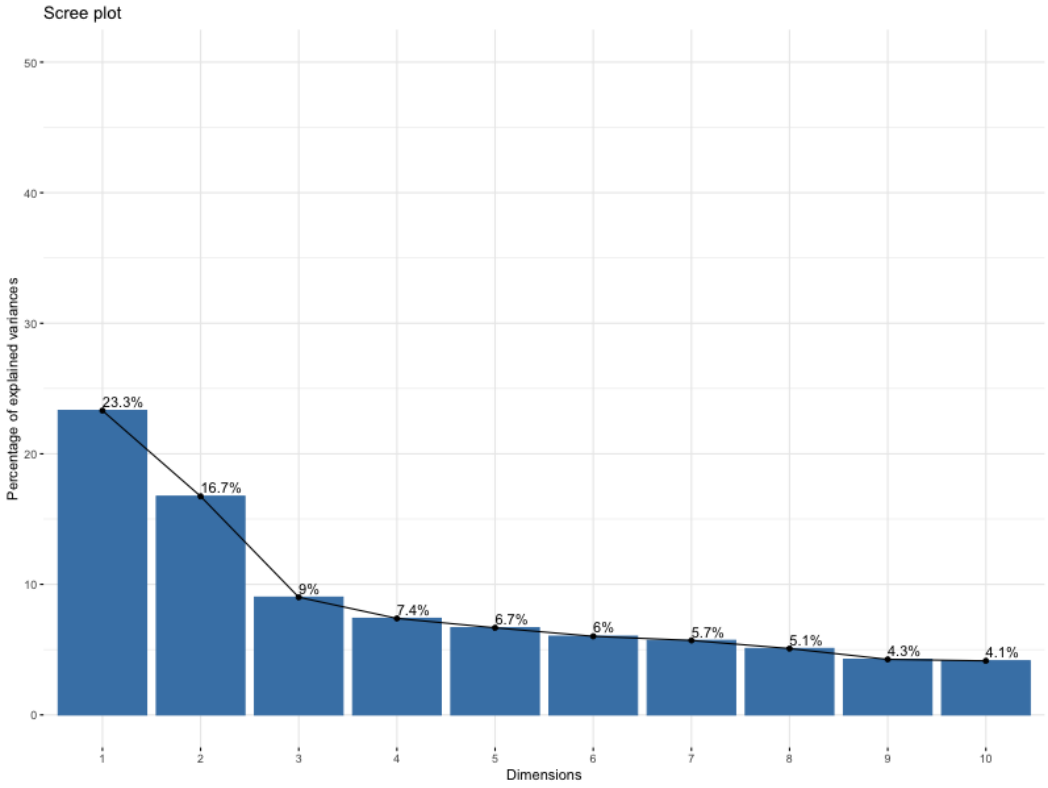


Figure 12: Scree plot showing the added explained variance with each principal component

```
null device
          1
Importance of components:
                          PC1    PC2     PC3     PC4     PC5     PC6     PC7     PC8     PC9
PC10    PC11    PC12
Standard deviation     1.9313 1.6368 1.20087 1.08747 1.03334 0.98156 0.95530 0.90183 0.82505
0.81404 0.71648 0.62748
Proportion of Variance 0.2331 0.1674 0.09013 0.07391 0.06674 0.06022 0.05704 0.05083 0.04254
0.04142 0.03208 0.02461
Cumulative Proportion  0.2331 0.4006 0.49069 0.56460 0.63134 0.69155 0.74859 0.79942 0.84197
0.88338 0.91547 0.94008
                          PC13   PC14    PC15    PC16
Standard deviation     0.54616 0.5138 0.50612 0.37461
Proportion of Variance 0.01864 0.0165 0.01601 0.00877
Cumulative Proportion  0.95872 0.9752 0.99123 1.00000
```

Figure 18: Summary of our PCA

Ostensibly, we observe one significant discrete jump in explained variance, which occurs after the second principal component. The explained variance by these two components is still low since we would be explaining less than 50% of the variance with them. We observed each the relationship of each PC with the original variables and obtain:

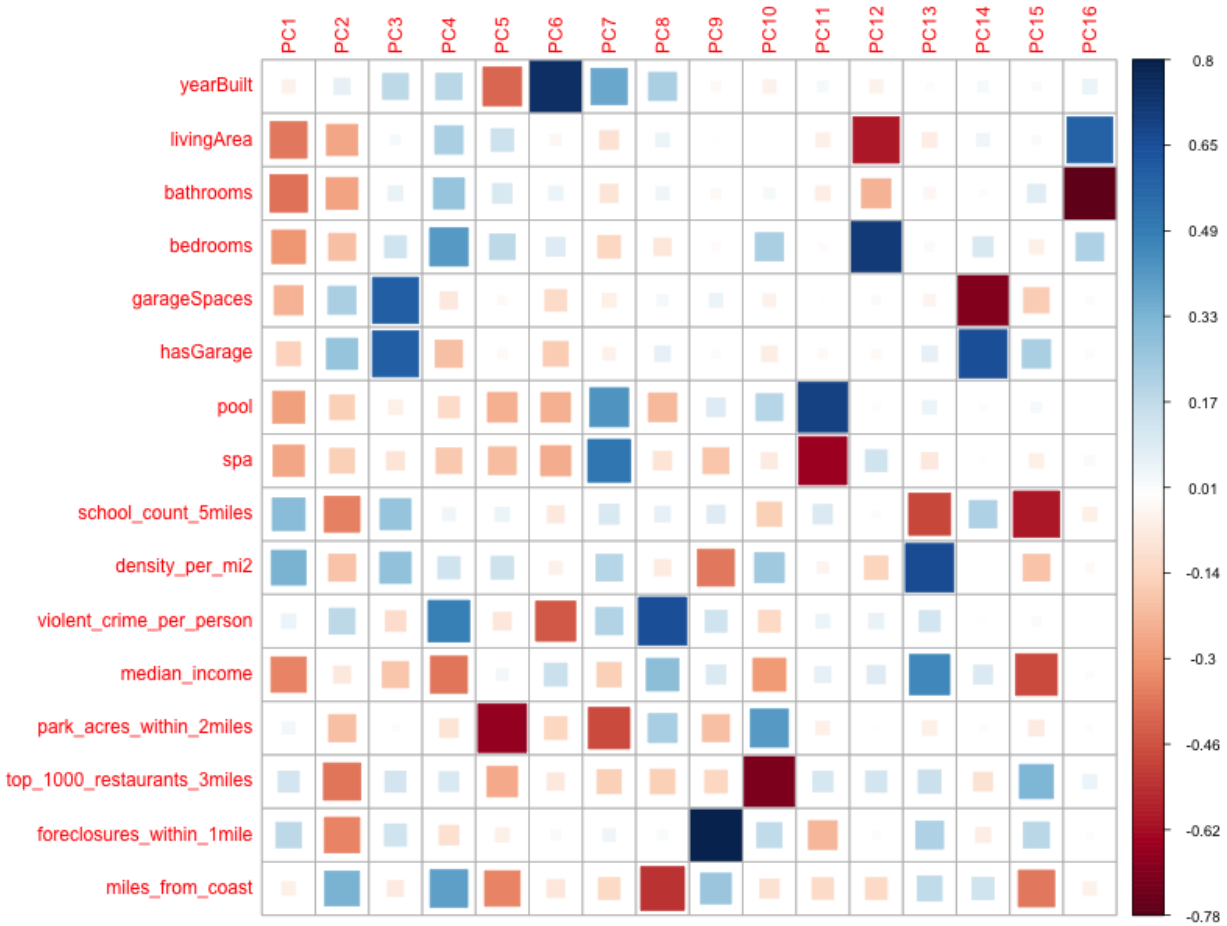


Figure 14: Correlation Matrix between our principal components and our features

The first principal component with the highest variance explanation relates to internal factors of the house and median income of the neighborhood. Other factors, such as city density and school count also affect this Principal Component. However, PC2 seems to concern these external factors even more closely, highly associating with miles from coast, foreclosures within one mile, highly rated restaurants, and school count as well. The analysis of how we use these principal components follows in the advanced analysis section, where we determine the

necessary principal components to predict housing prices.

4. Modeling Housing Prices

Initial Linear Model

We first performed an Ordinary Least Squares (OLS) regression, a method to estimate the parameters in a linear regression model. An OLS's purpose is to minimize the sum of the square errors, which are the differences between the observed dependent variable (house prices) and those predicted by the linear function.

Our initial OLS model analyzed logPPSF on all variables in our main LA data set and had an adjusted R-squared off of .60; however, with over 183 variables included, due to our categorical variables, dimension reduction through regularization and other methods is needed to get a more robust understanding of the factors that explain logPPSF in Los Angeles. This result may have occurred because, with a high dimensionality of variables, the high R-squared results from overfitting the data.

Cross Validated Lasso Regression

Cross-validation is a robust statistical method used to evaluate and compare the performance of predictive models by partitioning the original sample into a training set to train the model and a test set to evaluate it. This approach helps ensure the model generalizes well to new, unseen data.

We partioned our data into fifteen equally sized folds for our analysis and set our lambda ratio to .001. Additionally, we set our optimal lambda to use the 1se rule for causal inference.

Results:

The resulting cross-validation model reduced the number of variables present from 183 to 96 and received a resulting R-squared of .53. The initial OLS might have had a higher adj. R-squared of .60, however that model only saw internal data and thus is expected to have a higher R-squared than one trained on outside data.

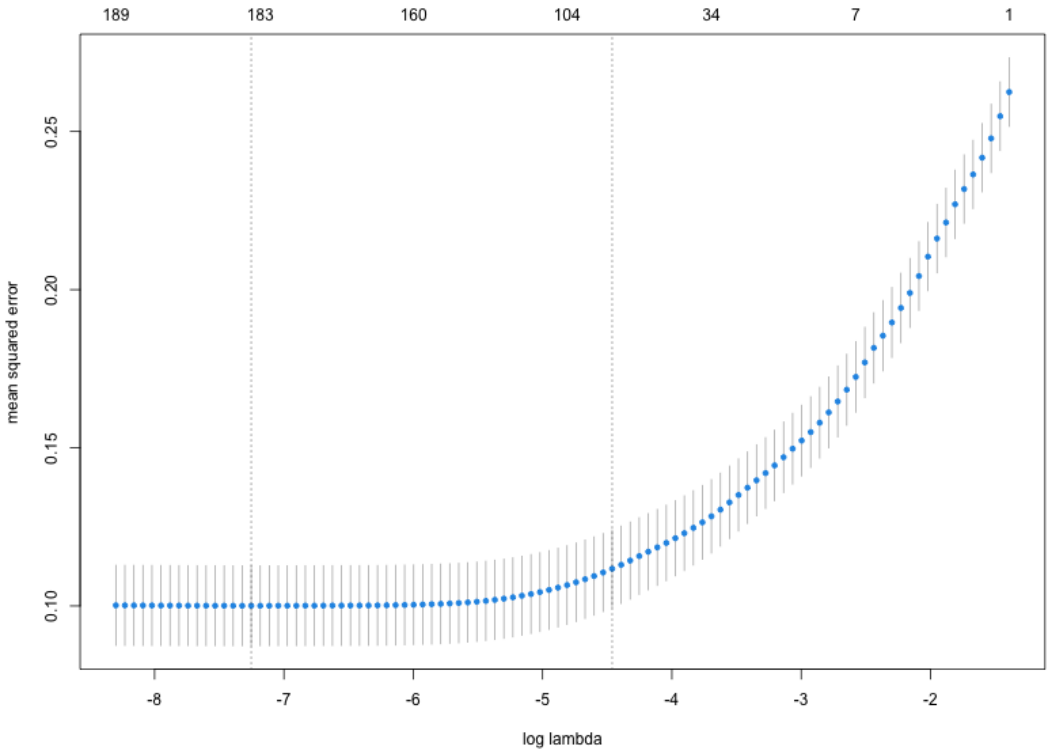


Figure 15: Cross-validation graph for our initial Cross-validated Lasso regression.

This graph shows our CV path with the X-axis representing the lambda (resistance) measure and the Y-axis representing the Out of Sample (OOS) MSE for each prediction. As stated above, our 1se level still has 96 variables and an MSE of about .12.

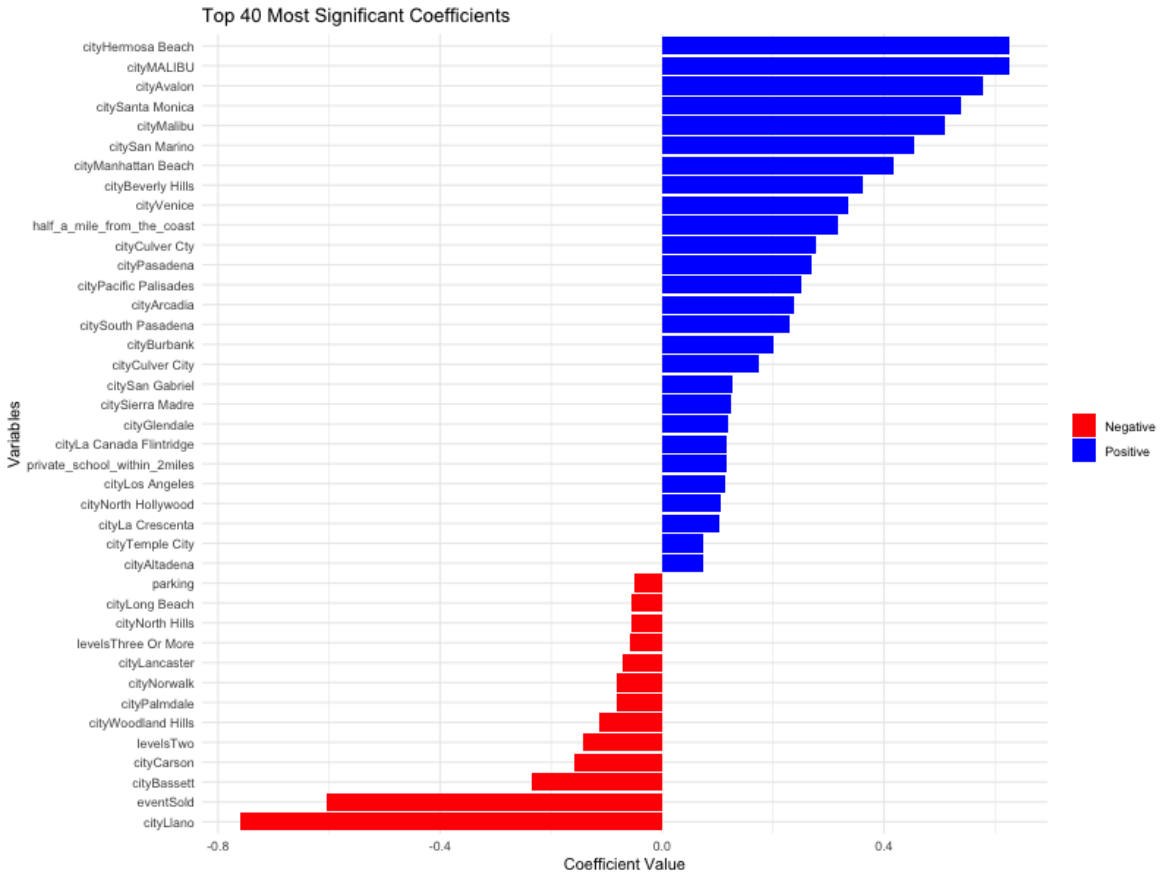


Figure 16: Most important coefficients in our initial Lasso model.

This graph shows the most significant variables following our Cross-validated Lasso regression. It finds that cities are the most dominant influence on housing prices, even after including neighborhood characteristics. This insight makes sense as the name of a city represents a heuristic for all the amenities included internally and externally in the house. However, these results also indicate that the geospatial variables we added do not make up for the information added by the city variable.

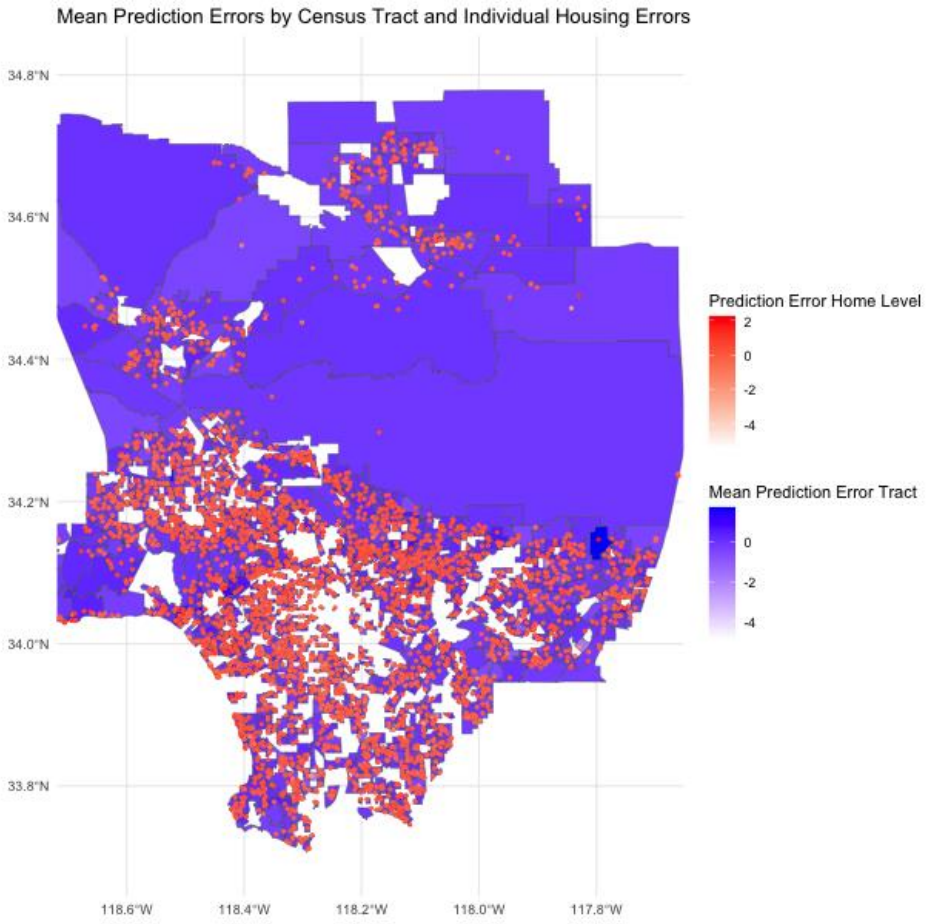


Figure 17: Map showing the spatial distribution of our errors.

While slightly saturated, this graph tracks the spatial correlation between our errors at the tract and individual house level. We find relatively uniform error rates among most tracts except those with few observations.

5. Advanced Analysis

Random Forest

Random Forest is an ensemble learning method that constructs multiple decision trees

during training and outputs the mean prediction of the individual trees. This approach helps avoid overfitting and improves the predictive power of the algorithm. For our analysis, we used Random Forest to predict the logPPSF of houses in LA county; we discarded categorical variables, like city, but kept all other data in the model. We split our dataset into training and testing sets, with 70% of our data being used for training and 30% for testing. Additionally, our Random Forest was trained on 500 trees.

Results:

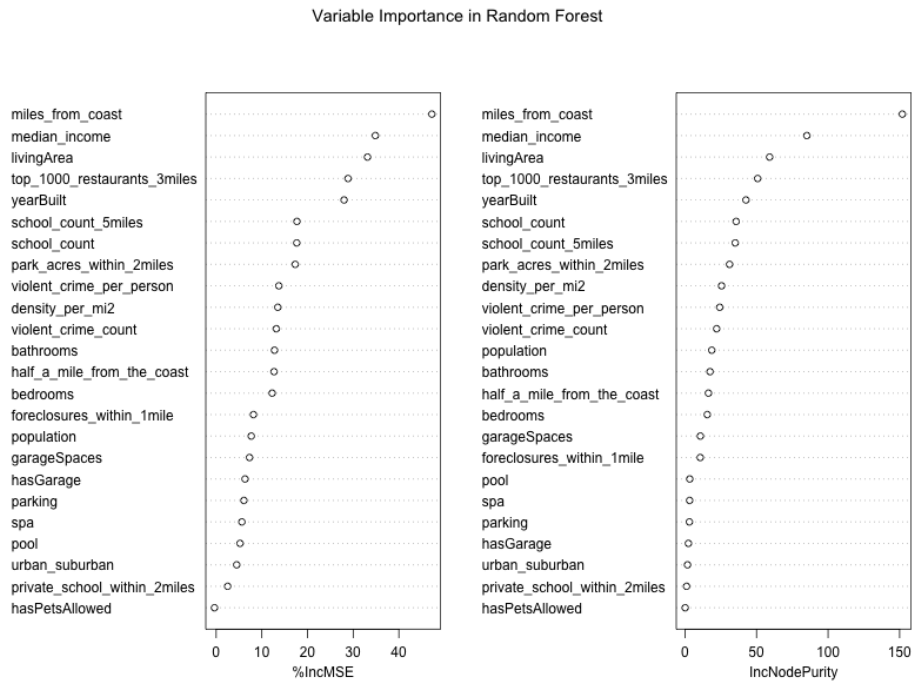


Figure 18: Variable importance table of our Random Forest Model

In our Random Forest, we were able to achieve an OOS R-squared of .64, improving upon our Cross Validated Lasso regression and an MSE of .09. We found that the most significant variable in this model was 'miles from the coast,' 'median income,' and 'living area' with their exclusion each adding over 30% to our MSE.

Additionally, we analyzed our logPPSF with gradient-boosted trees (XGBoost), which gave us an MSE of .10 and an OOS R-squared of .603. Collinearity, noise outliers, or suboptimal tuning could be the reason for the drop in performance moving from Random Forest to XGBoost.

CART

The Classification and Regression Trees algorithm, or CART for short, is another non-parametric decision tree learning technique that produces regression trees. CART works by splitting data into subsets based on feature values to create the most homogenous groups possible. For our analysis, we utilized CART to predict logPPSF. We maintained the same features for CART as we did for our Random Forest and XGBoost – removing categorical variables and splitting our data into 70% training and 30% testing data.

Results:

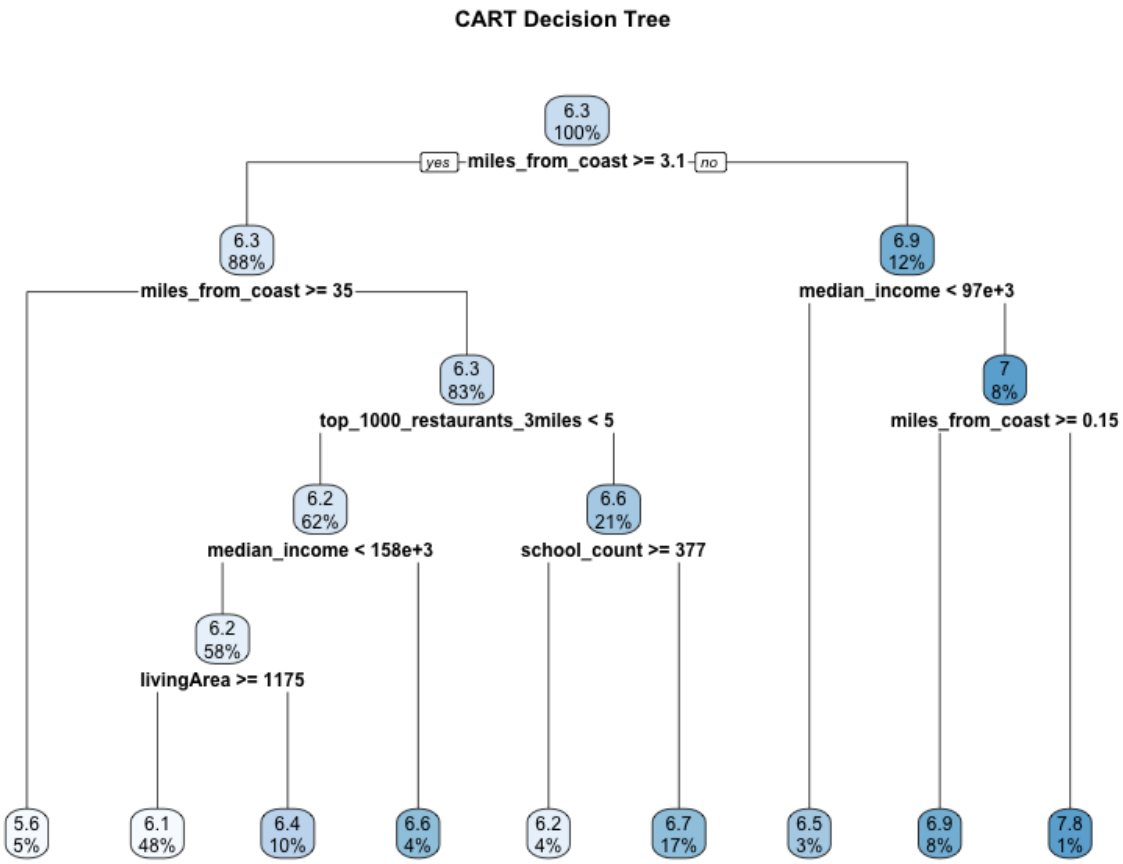


Figure 19: Visualization of our CART Decision Tree

The CART model achieved an OOS R-squared of .426 and an MSE of .147, indicating that it performed worse than our Random Forest, XGBoost, and Cross Validated Lasso regression.

This model shows that, like our Random Forest, distance from the coast is our most important feature, followed by median income and amenities like school count or number of top

1000 restaurants within three miles. We can read our CART tree by starting at the top and working our way down. For example, it classified homes worth more than 7.8 logPPSF as those that are closer than 3.1 miles from the coast, have a median area income of over $97,000, and are within .15 miles of the coast.

Principal Components Regression

Once identified the relevant principal components, we can conduct a LASSO regression to verify the relevance of these components to predict housing prices. Although we found a significant jump in explanation of variance, the lack of variance explained by each PC did not allow for a conclusive regression using a limited number of PCs. When use AICc and BIC to pinpoint an adequate number of PCs, we obtain a high number of PCs.

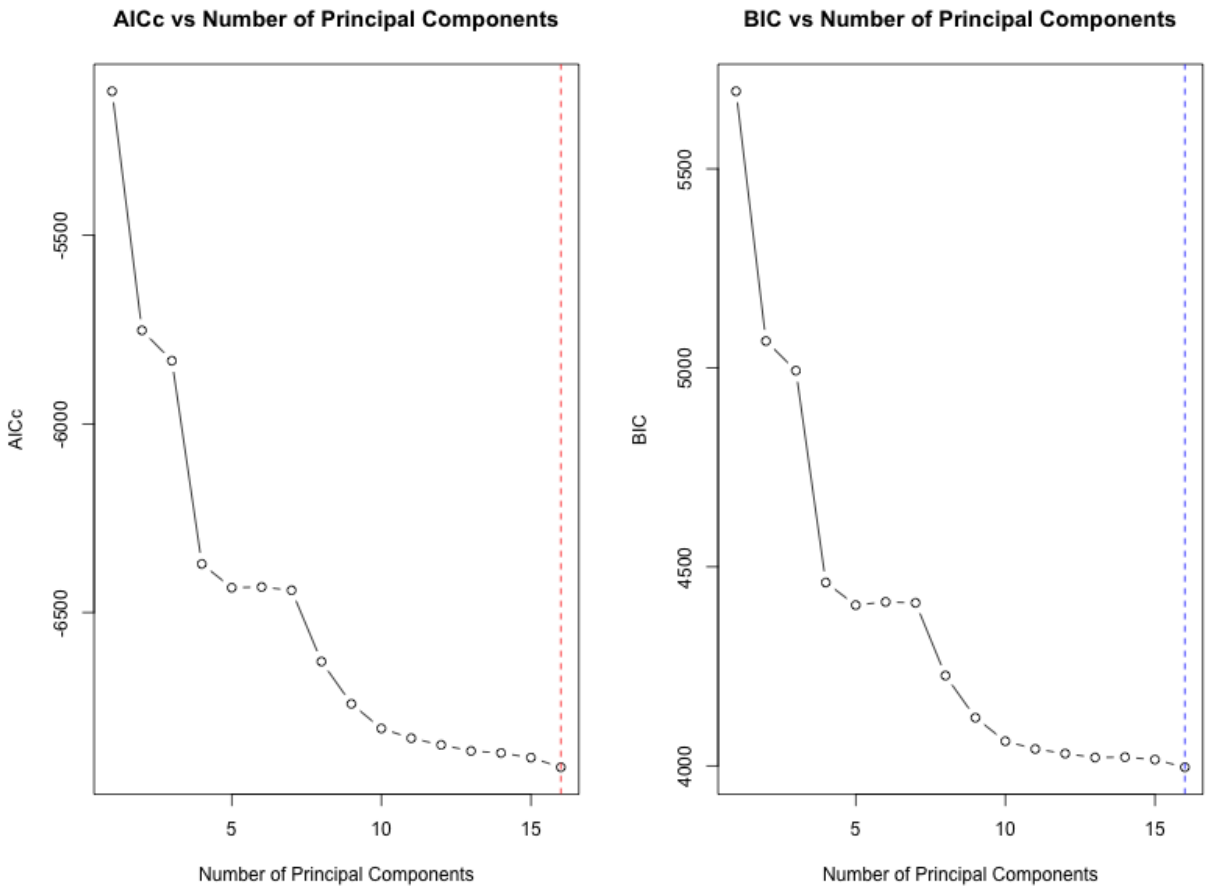


Figure 20: AICc and BIC test to choose what model is best for our PCR.

AICc's optimal number of principal components is 16, BIC also suggest 16. Moreover, the results of tunning a LASSO regression with the principal components yield the following table:

```
## 21 x 1 sparse Matrix of clas" "dgCMatr" x"
##                    s0
## (Intercept)  6.34074027
## PC1          .
## PC2         -0.10002304
## PC3          .
## PC4         -0.02820607
```

```
## PC5          0.10096628
## PC6         -0.06386615
## PC7          .
## PC8          .
## PC9          .
## PC10         .
## PC11        -0.00558025
## PC12         0.04236957
## PC13         .
## PC14         .
## PC15         .
## PC16         .
## PC17         .
## PC18         .
## PC19         .
## PC20         .
```

This table suggests that PC2, PC4, PC5, PC6, PC11, and PC12 are relevant in determining price variation. PC11 relates to luxurious internal factors that include pool and spa. Meanwhile, PC12 referred to number of bathrooms and bedrooms. PC5 and PC6 relate to external factors relating to nearby parks and crime rates (both with an association with year of construction), respectively. Overall, the following table shows the minimization of a cross-validated MSE using PCR:

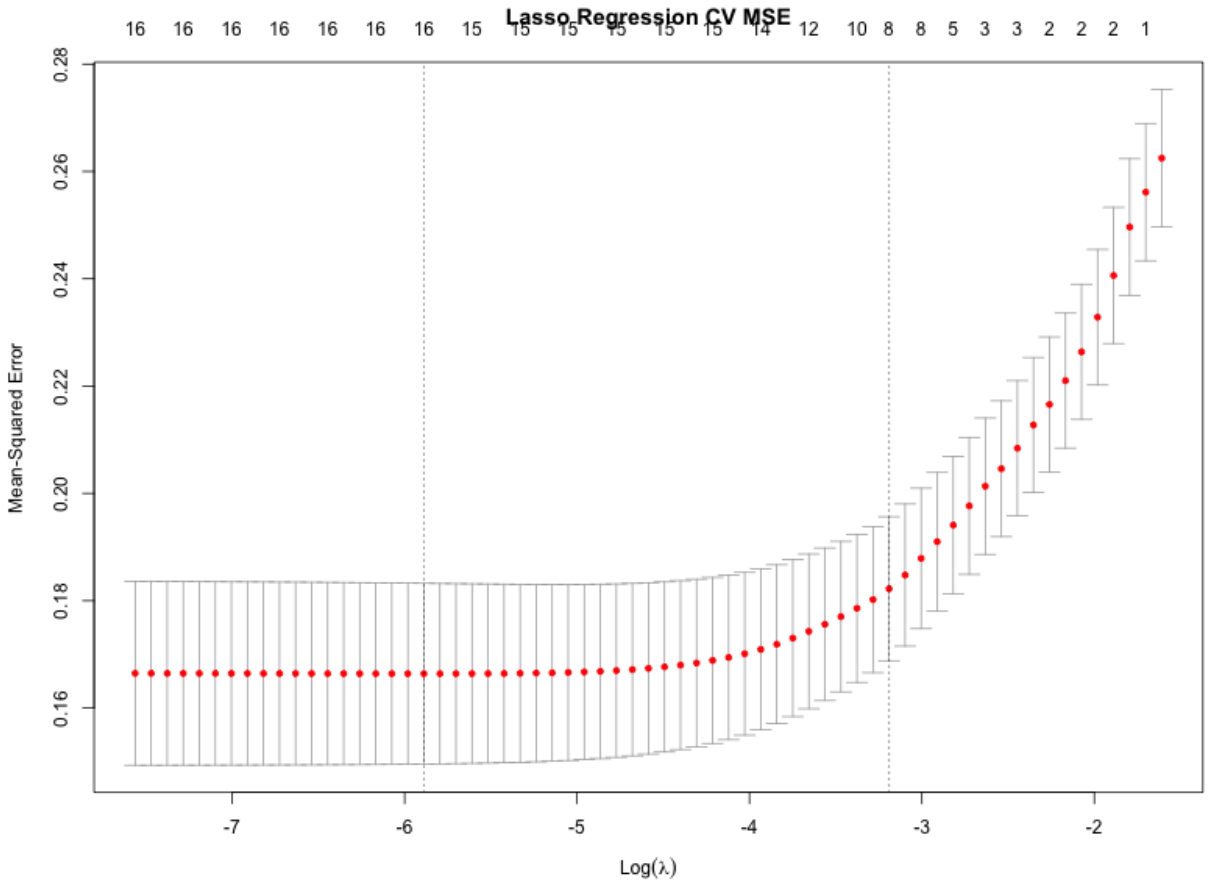


Figure 22: Cross-validated Lasso using our principal components.

Categorical Classification: Naïve Bayes

After classifying our data into quartiles, we test how well our model predicts the price quartile to which it belongs by associating new observations with other closely related observations already observed:

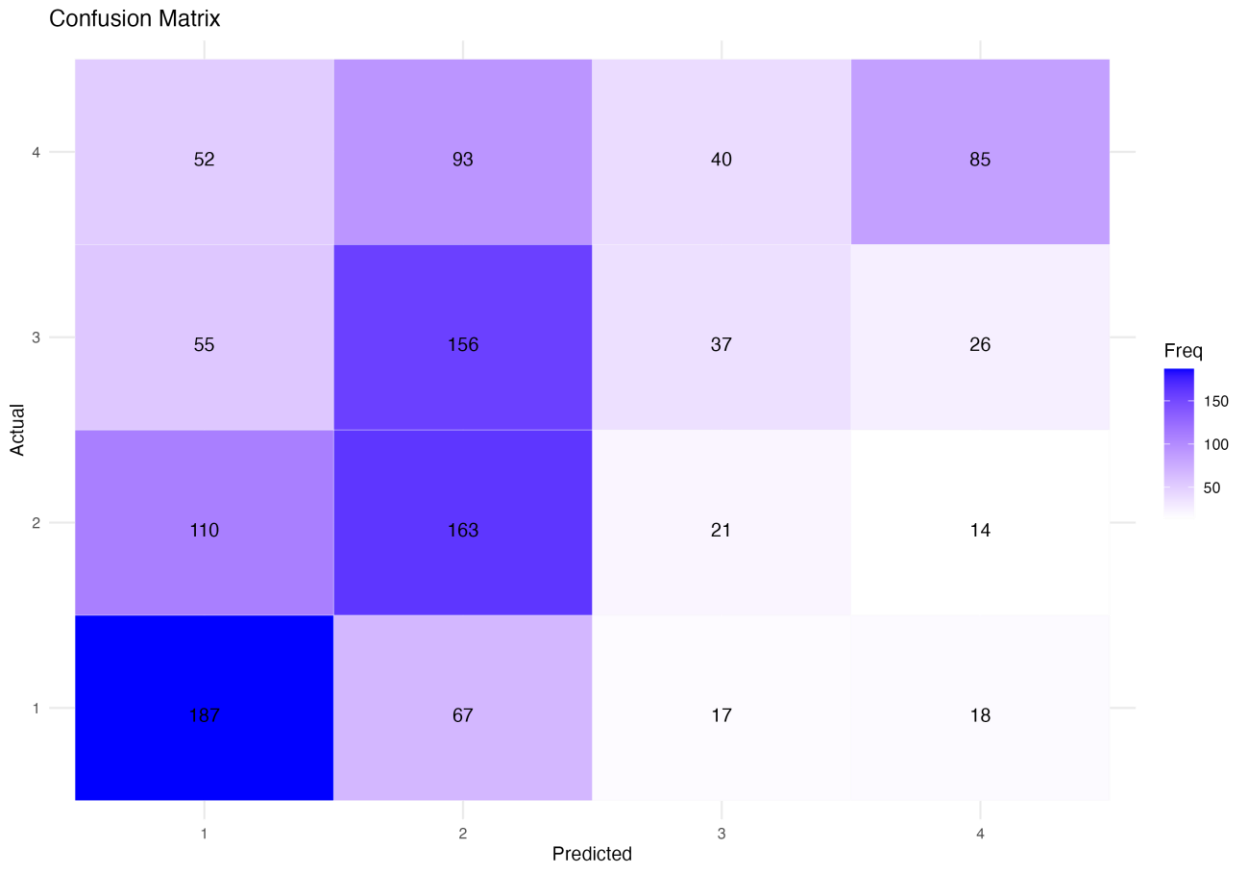


Figure 23: Spatial distribution of MSE from initial regression showing biases in Western suburbs and homes located near the coast.

The k-means performance of our model suggests high accuracies for predicting belonging to the first and fourth quartiles. Nonetheless, the performance is poor when classifying observations into the second and third groups.

6. Results and Comparison

Highest Model Performance:

We found that Random Forest gave us the best predictive performance for our housing model,

with an MSE of .09 and an OOS R-squared of .64.

Variable Importance:

We found that consistently, the most critical variable for classifying housing prices was distance from the coast, followed by median income. We found city names to have the most significant predictive power for models where categorical variables were included. Surprisingly, neither green space nor crime rate held strong predictive power for our tree-based models. The discrepancy in the predictive power of crime on housing prices could be due to geospatial merging issues (e.g., the NA values for the crime data) or it could be possible that housing only affects housing prices in extremes or in specific geographies. Explanations for the lack of predictive power of park space could be that rural areas generally have more green space regardless of neighborhood quality due to the surplus of open land, while urban areas generally have less park space due to the limitation of underutilized land.

Model Comparison:

We found that Random Forest had the best prediction power. XGBoost, Cross Validated Lasso regression, and CART followed. The table below shows the performance of these models:

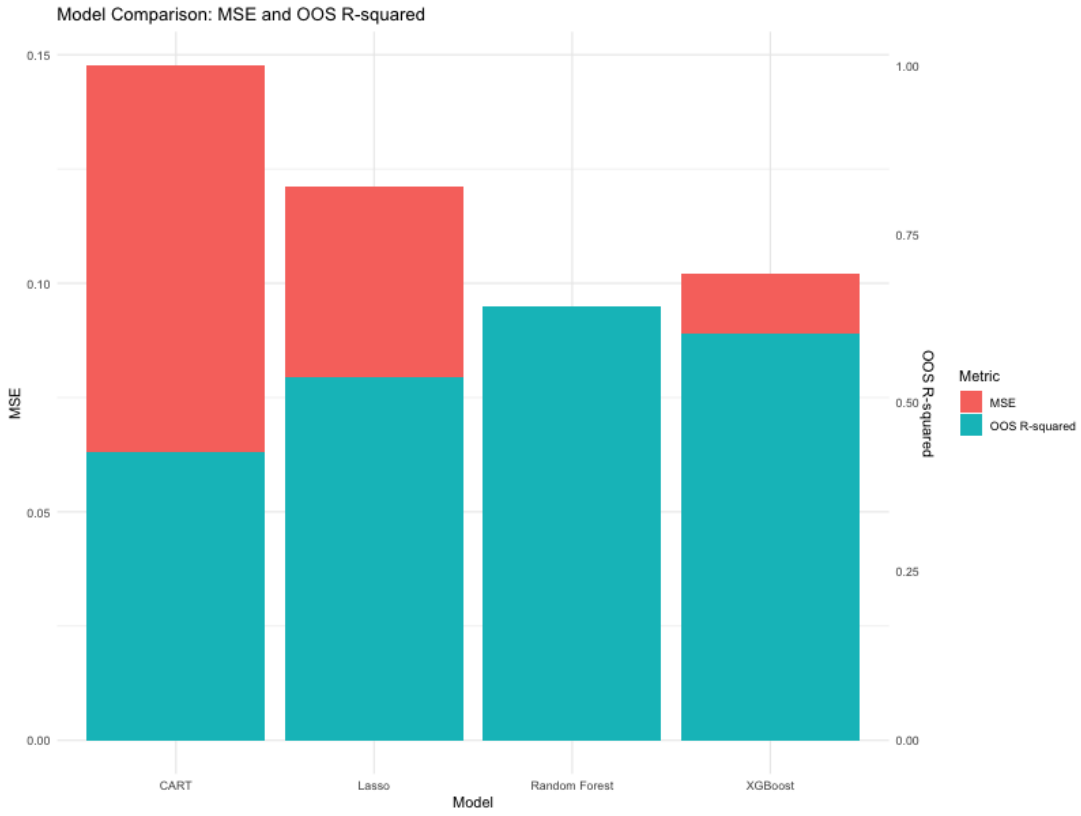


Figure 25: Out of Sample R-squared and Mean Squared Error for each model

Figure 13: Spatial distribution of MSE from initial regression showing biases in Western suburbs and homes located near the coast.

7. Conclusion

Our analysis found that external factors significantly impact housing prices and that we can explain over 60% of the variance in housing prices through our internal and external characteristics model. We also found that tree-based methods do better than OLS or Lasso regressions for predicting pricing in our model. Lastly, we learned that improvements can still be made to more accurately isolate the determinants of housing prices. Whether through adding

more features related to livability (e.g., walkability score, retail spaces, and nightlife) and neighborhood demographics or changing the level of geospatial analysis (e.g., removing census tracts and only using the radius or a more holistic measure), there's still many improvements to be made to increase its predictive power.

8. Future Work

Reflecting on the results and methodologies applied in this analysis, we identified several avenues for future research and enhancement of the model. The current study relies on cross-sectional data, limiting our ability to capture the temporal dynamics of housing prices. Future work should aim to incorporate panel data, which would provide a more robust analysis of how housing prices evolve and how temporal changes in variables such as economic conditions and neighborhood developments impact these prices.

Additionally, the analysis was constrained by the availability and completeness of data for certain census tracts. Not all census tracts in Los Angeles County were included in the study due to incomplete data, potentially leading to gaps in our understanding of the spatial distribution of housing prices and the influence of neighborhood characteristics. Furthermore, other relevant amenities, such as transportation, pollution, and technological levels, must be examined in addition to the ones examined herein. Future studies should strive for more comprehensive data coverage to ensure a holistic view of the housing market.

Expanding this work to other cities is another critical step. By applying the methodologies developed in this study to different urban areas, we can create an approach that accounts for

regional variations and identifies commonalities and differences in the factors driving housing prices. This comparative analysis across cities with varying characteristics, such as proximity to coastlines, availability of public amenities, and economic conditions, can help eliminate the idiosyncrasies specific to Los Angeles and uncover broader trends and determinants.

By addressing these areas, future studies can build on the findings of this research, providing deeper insights into the factors influencing housing prices and offering valuable guidance for policymakers, real estate investors, and homebuyers.